Properties, Prediction, and Prevalence of Useful User-Generated Comments for Descriptive Annotation of Social Media Objects

Elaheh Momeni
University of Vienna
Dept. of Computer Science
A-1090 Vienna, Austria
elaheh.momeni.roochi@univie.ac.at

Claire Cardie
Cornell University
Depts. of Computer Science
and Information Science
Ithaca, NY 14853, USA
cardie@cs.cornell.edu

Myle Ott
Cornell University
Dept. of Computer Science
Ithaca, NY 14853, USA
myleott@cs.cornell.edu

Abstract

User-generated comments in online social media have recently been gaining increasing attention as a viable source of general-purpose descriptive annotations for digital objects like photos or videos. Because users have different levels of expertise, however, the quality of their comments can vary from very useful to entirely useless. Our aim is to provide automated support for the curation of *useful* user-generated comments from public collections of digital objects. After constructing a crowd-sourced gold standard of USEFUL and NOT USEFUL comments, we use standard machine learning methods to develop a "usefulness" classifier, exploring the impact of surface-level, syntactic, semantic, and topic-based features in addition to extra-linguistic attributes of the author and his or her social media activity. We then adapt an existing model of prevalence detection that uses the learned classifier to investigate patterns in the commenting culture of two popular social media platforms. We find that the prevalence of USEFUL comments is platform-specific and is further influenced by the entity type of the media object being commented on (person, place, event), its time period (e.g., year of an event), and the degree of polarization among commenters.

Introduction

Descriptive annotations for social media objects by experts provide important supplemental information about the object (e.g., textual documents, images, videos) in the form of keywords and free-form descriptions. Usually comprehensive and of high quality, expert annotations are valuable both for human consumption and for aiding efficient information retrieval and resource management. But they are costly to create. User-generated comments, on the other hand, represent a potential complementary source of essential information like the names and places depicted in a photo or video — information that is often not available in existing metadata records (Ames and Naaman 2007; Kennedy et al. 2007).

For example, Flickr Commons allows libraries and museums to share their resources so that users can collaborate in the creation of descriptive annotations. One example of the results of this project is a photo (from the Library of Congress set) that was originally captioned simply as "Reid Funeral". It is now more fully described by the user-generated comment: "Photo shows the crowd gathered outside of the Cathedral of St. John the Divine during New York City funeral of Whitewall Reid, American Ambassador to Great Britain." [1]

Unfortunately, not all user-generated comments are useful. Users have different backgrounds, levels of expertise, and intentions for contributing comments. As a result, the quality of user-generated comments varies from very useful to entirely useless; comments can even be abusive or off-topic. And not surprisingly, what counts as a USEFUL comment can depend on a number of factors including the media type (e.g., document, video, art object, photo), the entity type of the object (e.g., is the object associated with a person, place, event), the time period associated with the object (e.g., early 20th century vs. the 1960's), or even the degree of controversy surrounding the object. Also important is whether usefulness is judged from the perspective of an institution, which might require objective and informative descriptive annotations, or from the perspective an end-user, who might value longer, more personal, or more subjective descriptions.

In spite of these complexities, methods for estimating the usefulness of user-generated comments are gaining increasing attention (Siersdorfer et al. 2010; Diakopoulos, De Choudhury, and Naaman 2012; Momeni and Sageder 2013). The most common approach simply allows all users to vote on (and possibly moderate) the contributions of others (Siersdorfer et al. 2010; Hsu, Khabiri, and Caverlee 2009; Lampe and Resnick 2004), thus avoiding an explicit definition of "useful". However, Liu et al. (2007) show that voting is influenced by a number of factors (e.g., a "rich get richer" phenomena) that distort accuracy.

The goal of the work reported here is to provide alternative, *automated* support for the curation of useful user-generated comments for use as descriptive annotations for digital objects. In addition, we aim to better understand the characteristics of useful user-generated comments and to estimate their prevalence across social media platforms. More

[1] Source: Library of Congress Flickr Pilot Project Report Summary, http://www.loc.gov/rr/print/flickr_report_final_summary.pdf.

specifically, we study two classes of digital object — photographs and videos — from two popular social media platforms — Flickr and YouTube, respectively.

We investigate usefulness from the user's perspective, defining a comment as USEFUL if it provides descriptive information about the object beyond the usually very short title accompanying it. With this definition in hand, we employ crowd-sourcing techniques to create a gold standard data set[2] of USEFUL and NOT USEFUL comments and propose the use of standard supervised machine learning techniques to develop a "usefulness" classifier that distinguishes useful from not useful user-generated comments. We consider over thirty features for the classifier including features for readability, informativeness/novelty, syntactic traits, named entity presence, sentiment, and topical traits of the text as well as features that describe the author's posting and social media behavior.

Our results are promising. We find first that the classifier identifies useful comments for Flickr photos with high reliability (precision (P) of 0.87 and recall (R) of 0.90), statistically significantly outperforming a strong baseline (P65, R80). Identification of useful comments on YouTube proves to be more difficult (P65, R83), but again the classifier statistically significantly outperforms the baseline (P55, R70).

Analysis of the top-ranked features of the classifier indicates that semantic and topic-based features are very important for accurate classification for both Flickr and YouTube, especially those that capture subjective tone, sentiment polarity and the existence of named entities. In particular, comments that mention named entities are more likely considered USEFUL; those that express the emotional and affective processes of the author are more likely NOT USEFUL. Similarly, terms indicating INSIGHT (e.g., think, know, consider) are associated with USEFULness while those indicating CERTAINTY (e.g., always, never) are associated with NOT USEFUL comments.

Next, we find that performance varies according to the entity type of the social media object. We look at three different entity types — people, places, and events — and find that the classifier has an easier time recognizing useful comments for people and events regardless of the social media platform. Training entity-type-specific "usefulness" classifiers generally allows improved performance over the type-neutral classifier results reported above.

Finally, we adapt an existing model of prevalence detection (Ott, Cardie, and Hancock 2012) that uses the learned usefulness classifier to investigate patterns in the commenting culture across social media platforms. We find different rates of useful comments for each platform with much higher rates for Flickr than YouTube regardless of the entity type of the social media object. Overall, we identify a general trend toward less useful comments as the time period associated with the object under discussion approaches present day; and a decrease in the prevalence of useful comments when polarization among the commenters w.r.t. the media object is higher. We believe that this is the first study to estimate the prevalence of useful user-generated comments for photographs and videos and the only study to date that aims to characterize useful comments for their descriptive annotation capabilities.

In the remainder of the paper we first describe related work and the creation of the gold standard corpus of user-generated comments. Next, we define the feature set used for constructing the "usefulness" classifier and present results when applying the classifier to the Flickr and YouTube data. Finally, we describe the approach for estimating the prevalence of useful comments on individual sites and examine the effect of social media platform, time period, and polarization on these estimates.

Background and Related Work

The related literature follows partially overlapping lines of research.

Assessing the usefulness of user-generated tags. Several works in the area of tagging and folksonomy research discuss the assessment of user-generated tags or the selection of tags that allow people to better describe their content. Sigurbjoernsson and van Zwol (2008) propose approaches for the selection of useful tags by computing tag and URL co-occurrence patterns. They find that the tag frequency distribution follows a perfect power law distribution, and indicate that the mid-section of this distribution contains the most interesting candidates for tag recommendation. Weinberger et al. (2008) define a measure of tag ambiguity, based on a weighted Kullback-Leibler (KL) divergence of tag distributions.

Assessing the quality of questions and answers. Agichtein et al. (2008) introduce a general graph-based classification framework for combining features from different sources of information in order to assess high-quality questions and answers in CQA (Community Question and Answer). Liu et al. (2008) propose a method for predicting information seeker satisfaction in CQA and develop a variety of content, structure, and community-focused features for this task. Harper et al. (2009) propose an algorithm that reliably categorizes questions as informational or conversational.

Assessing the quality of postings in micro-blogging services. Castillo et al. (2011) propose automatic methods for assessing the quality and credibility of a given set of tweets, first by analyzing postings related to trending topics, and then by classifying them as credible or non-credible. Diakopoulos et al. (2012) develop methods for filtering and assessing the variety of sources found through social media by journalists by using a human centered design approach. Becker et al. (2012) present relevant Twitter content selection approaches and show that the centroid (as a centrality-based approach) emerges as the preferred way to select relevant tweets given a cluster of messages related to an event.

Assessing the helpfulness of product reviews. Predicting the helpfulness of a product review (e.g., how many people have considered a particular product review helpful) is related to the problem studied here. Several approaches demonstrate that a few relatively straightforward features

[2]This dataset is available by request at http://homepage.univie.ac.at/elaheh.momeni.roochi/data-ugc

Platform		Event	Place	Person	Total
Flickr	Manual coding	1,200	1,100	1,200	3,500
Flickr	All	13,864	6,935	12,474	33,273
YouTube	Manual coding	1,500	2,000	1,500	5,000
YouTube	All	50,654	6,908	34,216	91,778

Table 1: Summary statistics for the dataset

Platform	Total	Useful	Not Useful	Agree
Flickr	3,500	1,345 (38.42%)	2,155 (61.57%)	0.86
YouTube	5,000	414 (8.28%)	4,586 (91.72%)	0.72
ALL	**8,500**	**1,759 (20.69%)**	**6,741 (79.30%)**	**0.79**

Table 2: Manual coding results across platforms. Agreement scores are assessed based on Mean Fleiss' Kappa scores.

can be used to predict with high accuracy whether a review will be deemed helpful or not. These features are length of the review (Kim et al. 2006; Ghose and Ipeirotis 2007), mixture of subjective and objective information (Ghose and Ipeirotis 2007), readability such as checking the number of spelling errors (Ghose and Ipeirotis 2007), and conformity (a review is evaluated as more helpful when its star rating is closer to the consensus star rating for the product) (Danescu-Niculescu-Mizil et al. 2009; Kim et al. 2006). Moreover, Lu et al. (2010) illustrate how social features of reviewers can help the assessment process. Although our task is different, we will rely on some of these features for the learning-based classifier.

Manual Coding and Data Collection

This section describes how we collect user judgments identifying useful comments from two social media platforms using a crowd-sourcing approach.

Datasets: We compiled a dataset from real-world comments harvested from the popular social media platforms, YouTube and Flickr. These provide free-text comments on media objects (video and photo) from a variety of people with different backgrounds and intentions. In order to analyze the correlation between usefulness and different attributes of media objects (entity type, time period, etc.), we first selected three broad entity types: *event*, *person*, and *place*. Second, we used the history timeline of the 20th century provided by About.com to identify topics associated with the selected entity types from each decade of the 20th century. The resulting topics included, among others, the "Irish civil war" and "1936 Olympics" as events,"old New York" and "old Edinburgh" as places, and "Neil Armstrong" and "Princess Diana" as people.

Next, we searched each of Flickr Commons and YouTube for photos/videos of each topic (if available), selecting those with the highest number of comments (Flickr) or with a high number of views and (at least 100) comments (YouTube). In total for Flickr we crawled 33,273 comments written on 11,102 photos. For YouTube we crawled 91,778 comments (the first $1,000$ for each topic) written for 310 different videos.[3] (Distribution of the comments across entity types is shown in Table 1.) For each comment from both platforms we crawled all profile information for the author, utilizing a language detection library[4] to identify English comments. As a result, we obtained comparable datasets from YouTube and Flickr for topics involving events, people, and places across different time periods starting in 1900.

[3]The list of topics is available by request at http://homepage.univie.ac.at/elaheh.momeni.roochi/data-ugc

[4]http://code.google.com/p/language-detection

Manual Coding for Usefulness. We randomly selected 3,500 comments from Flickr and 5,000 from YouTube for manual coding with respect to usefulness. (As will be seen below, more comments were required from YouTube due to the low rate of useful comments.) See Table 1 for the distribution of comments across entity types.

Annotators were obtained via the CrowdFlower.com crowd-sourcing platform, which distributed our task across different channels, such as Mechanical Turk or getPaid. We asked coders to assist us to define useful comments, showing each coder a comment and links to the related media object (Flickr photo or YouTube video). To ensure the quality of the work by coders, for each comment we asked the coder to answer three objective questions, the answers to which can be computed automatically, and a fourth question that addressed the usefulness of the comment. The first and second question for both platforms were semantically the same but asked in two different ways. Inconsistency in answering the first two questions gives us the chance to exclude randomly selected answers. The first two questions for the Flickr user study are: 1-"how many Web links does the comment contain?", 2- "does the comment contain Web links"? The first two questions for the YouTube user study are: 1- "Is the length of the video short or long?" (more than two minutes is long, less than two minutes is short) 2- "how long is the length of the video?" The third question required writing a text-based answer, offering an additional chance to exclude data from non-serious coders. The central question for the task was the following: *"Compared to the description provided by the uploader of the media object (located below the video or photo), is this comment useful for you to learn more about the content of the media object (video or photo)?"*. For each comment we collected three judgments.

The examples below show the range of comments judged by the annotators:

- USEFUL**: Flickr photo - Dr. F.A. Cook**[5]. "This must be Dr. Frederick A. Cook (1865-1940), the American explorer who claimed to have reached the North Pole in 1908, before Robert Peary. The controversy over his claim continues. Not only does he have a Wikipedia article, but there are websites dedicated both to disdaining him and to celebrating him. Old controversies never die; they just go on the Internet."
- NOT USEFUL**: Flickr photo - Capt. and crew of MACKAY-BENNETT**[6]. " My great grandfather was an engineer at that time. I'd love to get a list of the names in that photo."

[5]http://www.flickr.com/photos/library_of_congress/2850357813/comment72157607279573241

[6]http://www.flickr.com/photos/library_of_congress/2536790306/comment72157629444651496

- USEFUL**: YouTube video - Lady diana interview before wedding**[7]. "She had JUST turned 20 years old when they married-in fact it had been less than a month since her 20th birthday. She wasn't anything more than a teenager. So tell me- how good were you at judging character at that age eh?"
- NOT USEFUL**: YouTube video- World War I: Battle Of Verdun**[8]. "Rich people get their poor people to fight the other rich people's poor people. And the[n] we do it all over again. Humanity is truly retarded."

Results of Manual Coding. 1,759/8,500 comments (20.69%) received majority agreement on how useful they were. We assessed the level of the (inter-annotator) agreement among coders using Fleiss' Kappa. The mean Kappa score is above 0.79, indicating substantial agreement for the usefulness inference between coders. Table 2 shows detailed agreement statistics for each platform. Table 2 also shows that Flickr samples exhibit a much higher rate of useful comments than YouTube (38.42% vs. 8.28%), and the agreement on usefulness for YouTube is lower than that of Flickr (0.86 vs. 0.72).

Feature Engineering

As described in the Related Work Section, some relatively straightforward features and strategies derived from social media and textual content have been used in existing work to characterize with high accuracy whether user-generated content (Tags, Q&A postings, Tweets, and product reviews) is helpful, relevant, high quality, or credible. Therefore, we believe that similar classes of features will be useful in our social media context. We focus on features that can be extracted from both Flickr and YouTube; however, most are quite generic and can be applied to other platforms as well. Our feature set is listed in Table 3. We divide the features into three groups:

- *Text-based and Linguistic Features (***TL***):* This group captures surface-level identification of usefulness. It includes features that are based on aggregate statistics extracted from the text, such as the readability, informativeness, average sentence length, number of punctuation marks, number of different links, and part-of-speech (POS) tagging of the words in the comment. We collect statistics based on the POS tags to create features such as percentage of verbs, adverbs, punctuations, etc. We use POS taggers from the LingPipe toolkit [9].
- *Semantic and Topical Features (***ST***):* The meaning of a comment may increase or decrease its usefulness. This set includes features such as number of Named Entities, number of different types of Name Entities, subjectivity tone, sentiment polarity, and psychological characteristics of the content of comments. For features that rely on Named Entity recognition we used the GATE toolkit (gate.ac.uk). We used LIWC (Tausczik and Pennebaker 2010) to identify 80 classes of psychological dimensions

[7]http://www.youtube.com/watch?v=IRTuI37mua4
[8]http://www.youtube.com/watch?v=d2qamDMs-3g
[9]http://alias-i.com/lingpipe/

Features	Short Description
TL (Text-based and Linguistic Features)	
Readability	measures how difficult the comment is to parse using the Gunning fog index (Gunning 1952)
Informativeness	measures the novelty of terms, *t*, of a comment, *c*, compared to other comments on the same object, calculated using:$\Sigma_{t \in c} tfidf(t,c)$
Punctuation Mark	counts the number of punctuation marks
Text Statistics	measures aggregate statistics extracted from the text #Words, #Verbs, #Adverb, WPS (average length of sentences)
Linkage Variety	counts the number of unique hyperlinks in a comment
ST (Semantic and Topical Features)	
Named Entities	counts the number of named entities that are mentioned in a comment
NE Types Variety	counts distinct types of named entities (such as person, place, date, etc.) that are mentioned in a comment
Topical Conformity	measures the distance between the topics of a comment and the topics belonging to other comments on the same object. We use the Jensen-Shannon (JS) divergence to measure the topic distribution distance of all comments on an object (A) compared to the comment's topic distribution (C). $D_{JS} = \frac{1}{2}(D_{KL}(C \parallel A) + (D_{KL}(A \parallel C)$ and KL divergence is calculated as: $D_{KL}(C \parallel A) = \Sigma C(i) log \frac{C(i)}{A(i)}$.
Sentiment Polarity	measures the sentiment/polarity of a comment as: $SenPolarity = \frac{PositiveScore+NegativeScore}{\#Words}$ We use LIWC for identifying positive and negative scores.
Subjectivity Tone	measures the subjectivity degree of a comment. We use Subjectivity Lexicon (Wilson, Wiebe, and Hoffmann 2005) to calculate subjectivity
Author Topic Entropy	measures the topical focus of an author via the entropy of topic distributions of the author. We define entropy of topic distribution of all comments authored by an author, a_i as: $H(a_i) = -\Sigma_{j=1}^{n} p(t_{i,j}) \log p(t_{i,j})$, where t is a topic and n is #topics.
Psychological & Social characteristics of the content	identifies psychological dimensions: Leisure, Anger, Family, Friends, Humans, Anxiety, Sadness, Sexuality, Home, Religion, Relativity, Affective Process, and Self-reference scores (Tausczik and Pennebaker 2010)
AS (Author and Social Features)	
Author Linkage Behavior	counts the number of unique hyperlinks posted by a user. A high linkage balance indicates that linkage is part of the commenting behavior of a user.
Author Conversational Behavior	counts comments that contain a @reply
Author Activity	measures different activities completed by a user: #Comments (counts the number of comments authored by the user), #UploadedObjects (counts the number of media objects uploaded by the user), #Favorite Objects (counts the number of media objects selected as favorite by the user)
Author Social Relation	counts the number of contacts of the user and measures Prestige score (measures the number of the Flickr Commons members in the contact list of the user)

Table 3: Overview of Features

in the texts of comments including self-reference terms (e.g., usage of "I"), leisure terms (e.g., cook, chat, music), anger terms (e.g., hate, loathe), etc.

Furthermore, the ST features include standard topic-modeling features that measure the topical concentration of the author of a comment and topical distance of a comment compared to other comments on an object. We use LDA, Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003), to model topics separately for authors and comments and model author-specific topics by creating one "document" per author, containing all comments posted by that author. Similarly, we model media-specific topics by creating one "document" per media object, containing all the comments for that media object. For both models, we choose the following hyper-parameters: $\alpha = 50/T$, $\beta = 0.01, T = 1,000$.

- *Author and Social Features (***AS***).* The characteristics of authors and their social media activity may increase or decrease the likelihood of their comments being USEFUL. Due to the limited access to this type of information, this feature group includes light-weight features such as author linkage behavior, author conversational behavior, author activities (e.g., number of comments, posted by an author, number of uploaded objects), and author social behavior (e.g., number of contacts in contact list)

Usefulness Classifier

Following previous work (Momeni and Sageder 2013), here we describe the creation of the learning-based "usefulness" classifier, evaluate it on the manually coded comments, and analyze the impact of individual features for identifying useful comments.

Experimental Set Up

For training the usefulness classifier, we selected a balanced set of $1,000$ USEFUL comments and $1,000$ NOT USEFUL comments from the Flickr data; we selected 400 of each class from the YouTube data. For both platforms, only comments for which at least two out of three coders agreed were selected. Our experiments employ two classifier models — logistic regression (LR) and Naive Bayes (NB). Classifiers were trained using combinations of the feature subsets described above and were evaluated according to four measures: precision (P), recall (R), F1-measure (F1), and area under the Receiver Operator Curve (ROC). We designed two baseline approaches for comparison purposes:

Baseline1 predicts usefulness using only one feature, INFORMATIVENESS. This feature is demonstrated by Wagner et al. (2012) to be an influential feature for predicting the attention level of a posting in online forums.

Baseline2 predicts usefulness using only the SUBJECTIVITY TONE, which is a particularly strong baseline as a result of our feature analysis study.

Results of Evaluations of Different Classifiers. Classification results for the two baselines and various feature and classifier combinations are given in Table 4. The results demonstrate the effectiveness of using semantic and author-related features for inferring useful comments.

Features	Classifier	Flickr				YouTube			
		P	**R**	**F1**	**ROC**	**P**	**R**	**F1**	**ROC**
TL	LR	0.76	0.75	0.75	0.85	0.56	0.56	0.56	0.60
	NB	0.74	0.71	0.71	0.77	0.60	0.59	0.59	0.65
ST	LR	0.84	0.85	0.84	0.93	0.66	0.72	0.68	0.71
	NB	0.81	0.80	0.79	0.89	0.62	0.87	0.71	0.72
AS	LR	0.79	0.60	0.68	0.80	0.58	0.54	0.56	0.53
	NB	0.71	0.66	0.65	0.80	0.64	0.53	0.44	0.53
TL + ST	LR	0.85	0.85	0.85	0.89	0.68	0.72	0.70	0.72
	NB	0.79	0.79	0.79	0.88	0.63	0.84	0.72	0.72
ST+ AS	LR	0.85	0.85	0.85	0.93	0.67	0.66	0.67	0.71
	NB	0.84	0.83	0.83	0.92	0.61	0.81	0.70	0.69
TL+ AS	LR	0.84	0.83	0.83	0.90	0.62	0.67	0.64	0.67
	NB	0.80	0.77	0.77	0.86	0.61	0.87	0.71	0.72
ALL	**LR**	0.87	0.90	**0.89**	**0.94**	0.66	0.74	0.70	0.72
	NB	0.84	0.83	0.83	0.91	0.65	0.83	**0.73**	**0.72**
Baseline1	LR	0.61	0.53	0.57	0.59	0.51	0.50	0.50	0.52
Baseline2	LR	0.65	0.80	0.72	0.77	0.55	0.70	0.61	0.59

Table 4: Results from the evaluation of classification algorithms with different feature settings (**bold** indicates the top F1 and ROC scores for each dataset)

In particular, in both datasets, training a classification model using author and semantic feature shows improved performance compared to the same models trained using text features. In the case of the Flickr dataset, we are able to achieve an F1 score of 0.89, coupled with high precision and recall, when using the Logistic regression classifier in combination with all features.

However, we find a lower level of F1 score (0.70) when using the same classifier on the YouTube dataset. For YouTube we are able to achieve an F1 score of 0.73 when using the Naive Bayes classifier. ROC measures show similar levels of performance for each classifier over the two sets.

In order to generalize the results of our evaluations, we analyze the diversity between the prediction results of the two best performing classifiers on YouTube: we apply Pearson's Chi-squared test ($p < 0.45$, $X2 = 0.045$), which indicates that there is no significant difference between them. Thus, this experiment identifies the Logistic Regression classifier using all features as the best-performing model w.r.t. F1 score for both platforms.

Influence of Features on Usefulness Classifier. So far we have only analyzed the use of various groups of features. Here, we will evaluate the quality of individual features for inferring the usefulness of comments for each dataset.

To determine how the features were associated with comment usefulness, we inspect the coefficients of the best-performing logistic regression model (using all sets of features). Positive feature weights correspond to the positive class (USEFUL), while negative weights correspond to the negative class NOT USEFUL). In addition to interpreting the statistically significant coefficients we also ranked the best performing features according to their Information Gain Ratio (IGR). Table 5 gives coefficients for the top-ranked features.

The top-ranked features from each dataset are dominated by Semantic and Topical features. Figure 1 shows the con-

Rank	Flickr					YouTube				
	Feature	All	Place	Person	Event	Feature	All	Place	Person	Event
1	**ST-Subjectivity Tone**	**-3.828**	-4.271	-6.228	-3.406	**ST-Subjectivity Tone**	**-1.499**	-0.129	-2.386	-2.002
2	**ST-Sentiment Polarity**	**-1.157**	-0.157	-0.223	-0.647	**ST-#Name Entities**	**0.157**	0.049	0.124	0.209
3	**ST-NE Types Variety**	**0.550**	-0.138	0.113	0.776	**ST-Self-reference**	**-0.126**	-0.148	-0.46	-0.360
4	**AS-Author Linkage Behavior**	**0.025**	0.046	0.003	0.002	**ST-Swear**	**-0.167**	-0.002	-0.571	-0.145
5	**ST-#Name Entities**	**0.211**	0.203	0.109	0.201	**ST-Sentiment Polarity**	**-0.014**	-0.023	-59.734	-0.173
6	**ST-Self-reference**	**-0.148**	-0.161	-0.136	-0.177	**ST-NE Types Variety**	**0.042**	-0.109	-0.175	0.328
7	**ST-Author Topic Entropy**	**-0.049**	-0.112	-0.302	-0.059	**ST-Anger**	**0.055**	-0.188	-0.138	-0.131
8	**ST-Insight**	**0.049**	-0.124	0.081	0.064	**ST-Tentative**	**0.051**	0.171	0.051	0.120
9	**ST-Swear**	**-0.045**	-0.005	-90.427	-3.363	**AS-#UploadedObject**	**0.084**	0.015	1.556	0.014
10	**TL-Linkage**	**0.173**	0.084	3.028	0.610	**TL-Future Verb**	**-0.143**	-0.426	-0.182	-0.298
11	**AS-Author Conversational**	**-0.023**	-0.086	-0.086	-0.066	**ST-Certainty**	**-0.012**	0.023	-0.034	-0.003
12	**ST-Certainty**	**-0.032**	0.110	0.042	-0.054	**AS-Author Conversational**	**0.027**	-0.154	-0.484	0.083
13	**TL-Future Verb**	**-0.043**	-0.071	-0.027	-0.027	**ST-Anxiety**	**-0.134**	-0.216	-0.339	0.008
14	**TL-Impersonal-pronoun**	**0.025**	-0.052	-0.040	-0.042	**TL-Impersonal-pronoun**	**-0.013**	-0.018	0.041	-0.087
15	**AS-Prestige score**	**0.060**	0.162	0.005	0.070	**ST-Friend**	**-0.032**	-0.519	-0.046	-0.011
16	**ST-Religion**	**0.089**	0.361	0.322	0.089	**ST-Religion**	**0.016**	0.046	-0.017	0.021
17	**ST-Sadness**	**-0.075**	-0.110	-0.403	-0.038	**ST-Sadness**	**0.036**	0.325	-0.218	0.289
18	**ST-Sexual**	**-0.014**	-1.306	-0.812	-0.284	**ST-Sexual**	**-0.059**	-0.007	-0.175	-0.059
19	**ST-Family**	**0.016**	-0.196	1.111	-0.004	**ST-Home**	**-0.355**	-1.760	0.692	-0.611
20	**ST-Relativity**	**-0.006**	0.163	-0.160	0.029	**ST-Family**	**-0.019**	-0.233	0.352	0.031

Table 5: Top-20 features for each platform and related coefficient ranks derived from the Logistic Regression model. Features are ranked based on Information Gain Ratio.

tributions by each of the top-5 features, where the affective process (such as Subjectivity Tone and Sentiment Polarity) and named entity-related features of the comments appear to play important roles for inferring useful comments for both platforms.

More precisely, coefficient ranks show that comments that express emotional and affective processes of the author (higher *Subjectivity Tone*, *Sentiment Polarity*, *Anger*, *Sadness*, *Swear*, and *Anxiety* scores) are more likely to be inferred as NOT USEFUL. *Subjectivity Tone* is a very good indicator for both platforms. Higher *Subjectivity Tone* has negative impact on the usefulness classifier. Furthermore, comments with offensive language (higher *Swear* score) are more likely to be inferred as NOT USEFUL. An analysis of the *Swear* and *Anger* scores between different platforms shows that YouTube contains more offensive language. Therefore, the *Swear* and *Anger* scores for YouTube are more negative than the Flickr swear score. However, these ranks show that comments that have higher *#Named Entities*, *NE Type Variety* and *Linkage* scores contain potentially interesting information and are likely to be inferred as USEFUL.

Usage of terms in LIWC's `insight` category (such as think, know, consider) shows good correlation with usefulness. Furthermore, terms in LIWC's `certainty` category (such as always, never) has a negative impact on the model. This might be due to the fact that authors who are assertive and express certainty tend to be seen as more subjective and less analytical. In contrast, using terms in LIWC's `tentative` category (such as maybe, perhaps, guess) shows that authors do not make any claims as to the correctness or certainty of their comments and such comments are likely to be determined USEFUL. Interestingly, *Readability* features are assigned little weight by the classifier. We suspect that this is because, while comments that are longer and contain more complex words are less "readable" based on the Gunning fog score, such comments are not necessarily less useful than comparatively shorter or less complex comments.

With regard to Author & Social features, *Author Linkage Behavior* is a good indicator showing that authors may diligently cite references for the information they provide. This increases reliability when inferring such comments as USEFUL. Similarly, we note that a higher *Linkage* score has

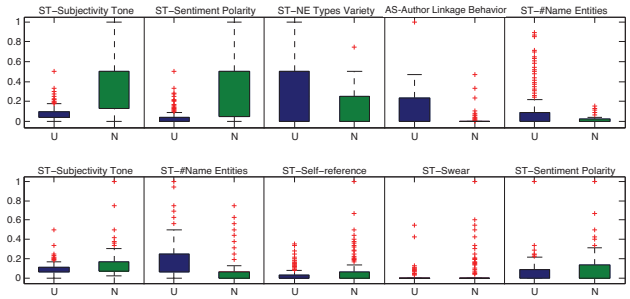

Figure 1: Top 5 Features for Useful(U), Not Useful(N) comments; Flickr(top), YouTube(bottom) (Momeni and Sageder 2013)

a positive impact on the usefulness inference, which is in line with the correlation of User Linkage Behavior score. A higher score of *Self-reference* and a higher *Author Conversational* score also have a negative impact. This suggests that authors who mostly use systems to converse and describe their personal experiences do not write useful comments. Interestingly, a higher *Author Topical Entropy* score of authors has a negative impact on the usefulness inference. This indicates that authors with a higher entropy have a lower topical focus and therefore write a comment with a lower level of focus and knowledge about the specific topic. Therefore, their comments are likely to be inferred as NOT USEFUL. Furthermore, for Flickr we note a higher *Contact* score does not have a negative impact. However, a *Prestige* score has a positive impact. This indicates that having influential contacts in the contact list is more important than having a higher number of contacts.

Influence of Entity Type of Topic on Classification. In all reported results so far, we have largely ignored differences due to the entity type being discussed. To explore the effects of entity type of topics on classifying a comment's usefulness, we divide the data according to the three types (person, place or event) being discussed. For each type, we then compare the performance of two classifiers: a *type-specific classifier*, which we train using only data of the same type as the test set, and an *type-neutral classifier*, which we train using data from all three types.

The results for type-specific and type-neutral classifiers

Platform		Person		Place		Event	
		All	Person	All	Place	All	Event
Flickr	F1	0.82	0.89 *	0.73	0.87 *	0.93	0.94
	ROC	0.93	0.97	0.93	0.97	0.96	0.96
YouTube	F1	0.70	0.80 *	0.67	0.74	0.82	0.84 *
	ROC	0.74	0.89	0.75	0.83	0.85	0.88

Table 6: Results from the evaluation of usefulness classifiers for different entity types. *All* is the type-neutral classifier, which is trained on data corresponding to all types of topics. * indicates a significant difference ($p < 0.01$).

are given in Table 6. We find that, in general, performance is better when the classifier is trained on comments of a single type, i.e., the classifier is type-specific, whereas performance is worse when the type is ignored, i.e., the classifier is type-neutral. We additionally perform three Pearson's Chi-squared tests between the prediction results of each classifier for each entity type. In Table 6, * indicates a significant difference at a $p < 0.01$ level for some types.

Furthermore, we investigate the importance of each feature for each topic with regard to usefulness inference. Table 5 shows detailed coefficient ranks for different models. Our discussion of the results focuses on the difference between the classifiers derived for each of the topics. An analysis of the most important features among different entity types of topics (place, person, and event) shows some differences. The major differences appear among the psychological characteristics of the content, but a few differences appear among other semantic and user features. There is no significant difference among text features.

More precisely, coefficient ranks show that comments related to the topics person and event express the author's emotional and affective processes more. These contribute to a comment being classified as NOT USEFUL. An analysis of the *Subjectivity Tone* among different topics shows that the *Subjectivity Tone* for topics related to person-related topics is higher than for other topics. An analysis of the *Swear* score among different topics shows that the *Swear* score for topics related to person is the most negative one. With regard to the topics related to event, the *Swear* score is more negative than for topics related to place. For topics related to person, *Family*, *Health* and *Body* scores have a positive impact on the model. This might be due to the fact that people describe more about various health and bodily aspects of a person on these topics. Furthermore, they describe the background of family members of the target person. This information may be useful information for other people. It is interesting to note that for the topic related to place *Relativity* scores have a positive impact on the model. However, *Friend* and *Family* scores have a negative impact on the model. This might be due to the fact that people describe more various physical phenomena and motion processes on this topic, which may be seen as useful information by others. Instead, giving information about friends and family is NOT USEFUL for others. With regard to topics related to event, event is a topic which often unifies place and person topics. This means that a topic related to event is often also related to person, place or both. Therefore, the coefficient ranks are influenced by the two other topics. For example, the *Relativity* score which includes physical place and motion has a positive impact on place and event, while it has a negative impact for person.

Our results indicate that there are a few relatively straightforward features that can be used to infer the usefulness of comments. However, an analysis of the important features across different platforms and different entity types reveals that when inferring usefulness, the impact of features varies slightly. The major differences appear among the psychological and social features (derived from LIWC) of the content. Therefore, a classification model should be trained that takes into account the topic of media objects for a more accurate classification of useful comments.

Prevalence of Useful Comments

This section aims to understand patterns in authors' comments peculiar to a particular commenting culture on different platforms and different dimensions (entity type, time period, and polarization) of topics of media objects. For estimating the prevalence of useful comments we adapt an existing Bayesian Prevalence Model (Ott, Cardie, and Hancock 2012) that uses the learned usefulness classifiers (see Table 6). The Bayesian Prevalence Model estimates the prevalence of useful comments in a set of comments by correcting the output of the noisy usefulness classifiers based on the performance characteristics of the classifiers. In the following section, first, we describe the formal definition and usage of the Bayesian Prevalence Model in our scenario and then we describe our experimental set up for estimating the prevalence of useful comments.

Bayesian Prevalence Model

Given an imperfect usefulness classifier, f, and a set of unlabeled comments, C_U, our goal is to use f to estimate the rate, or prevalence, of useful commenting in C_U. This task is challenging since f can produce both false positive and false negative predictions, and, therefore, cannot be relied on directly. Furthermore, if the probability of a false positive is different from the probability of a false negative, then the error introduced by f will vary depending on the true rate of useful commenting in C_U.

To address these challenges, we adopt the Bayesian Prevalence Model, introduced by Ott et al. (2012) to estimate the prevalence of deceptive online reviews, and jointly model our classifier's false positive and false negative rates, as well as the true rate of useful commenting in C_U. Formally, let us define our classifier, $f : \mathbf{c} \rightarrow y$, as a function mapping a comment, $\mathbf{c} \in \mathbb{R}^{|V|}$, to a usefulness label, $y \in \{0, 1\}$, where $|V|$ corresponds to the number of features. We further define f's *sensitivity* (true positive rate), η^*, and *specificity* (true negative rate), θ^*, as:

$$\text{sensitivity} = \eta^* = \Pr(f(\mathbf{c}) = 1 \,|\, y = 1),$$
$$\text{specificity} = \theta^* = \Pr(f(\mathbf{c}) = 0 \,|\, y = 0).$$

Then, in order to estimate the true rate of useful commenting in C_U, π^*, we model the process by which f makes its predictions. In particular, we model predictions made by f as a generative process with the following storyline:

- Sample the rate of useful commenting: $\pi^* \sim \text{Beta}(\boldsymbol{\alpha})$
- Sample the classifier's sensitivity: $\eta^* \sim \text{Beta}(\boldsymbol{\beta})$
- Sample the classifier's specificity: $\theta^* \sim \text{Beta}(\boldsymbol{\gamma})$
- For each comment, $\mathbf{c}$, in C_U:
  - Sample the comment's usefulness: $y \sim \text{Bernoulli}(\pi^*)$
  - Sample the classifier's prediction:

$$f(\mathbf{c}) \sim \begin{cases} \text{Bernoulli}(\eta^*) & \text{if } y = 1 \\ \text{Bernoulli}(1-\theta^*) & \text{if } y = 0 \end{cases}$$

Following Ott et al. (2012), we treat η^* and θ^* as latent variables with prior probabilities, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, set based on the cross-validation results in the previous section (see Table 6). We perform inference for this model with 70,000 iterations of Gibbs sampling, with 20,000 burn-in iterations and a sampling lag of 50. See Ott et al. (2012) for sampling equations and full derivation details.

Experimental Set Up

We set up three different experiments. First, for exploring the influence of time periods of topics on usefulness prevalence, we create 10 sets of comments related to each decade of the 20th century. Second, to explore the influence of a topic's polarization on its usefulness prevalence, we create 10 sets of comments from topics with varying degrees of polarization. Third, for exploring the influence of entity types of topics on usefulness prevalence, we create 6 sets related to each platform, that is for each platform one set for each entity type of topic (person, place and event), in total 26 sets. For each set of each experiment we used learned usefulness classifiers (see Table 6) and we predicted the usefulness of each comment and then we instantiated the Bayesian Prevalence Model in order to estimate the realistic rate of the different sets of comments related to the different dimensions of topics.

Influence of Time Periods of Topics on Usefulness Prevalence. In order to observe the effects of the time period of the topics (e.g, year of an event) on the prevalence of useful comments, we explore the prevalence for useful comments among different time related sets of comments, which belong to different time periods (different decades of the 20th century). Our results (shown in Figure 2) demonstrate that the temporal dimension of topics has slight influence on the usefulness prevalence. The nearer the time period of a topic is to the present time, the lower the prevalence of useful comments is. This might be due to the fact that topics related to earlier periods are less relevant to present time, therefore authors express less emotion and give more objective information, which may be inferred as useful information.

Influence of Polarization Degree of Topics on Usefulness Prevalence. Our next experiment explored the relationship between the prevalence of useful comments and the polarization degree of topics of media objects. Following Siersdorfer et al. (2010), by "polarizing topic" we mean a topic likely to trigger diverse sentiments and opinions among commenters, such as topics related to a presidential election in contrast to rather "neutral" topics such as

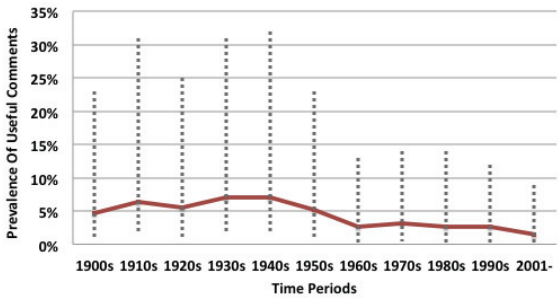

Figure 2: Graph of Bayesian estimates of usefulness prevalence versus time periods and polarization of topics. Error bars show Bayesian 95% credible intervals.

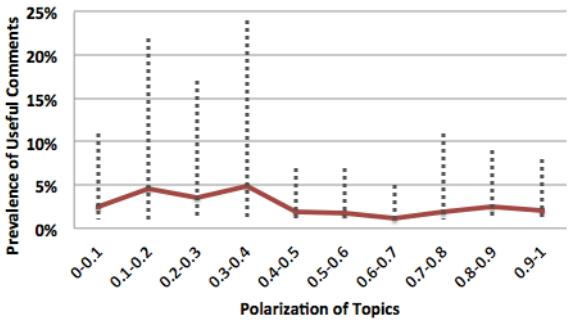

Figure 3: Graph of Bayesian estimates of usefulness prevalence versus polarization of topics. "0" shows that the topic of the video is not polarized while "1" shows the highest polarization.

"Ford Introduces the Model-T". In order to assess the polarization degree of topics we leverage the results of an exciting study (Siersdorfer et al. 2010) on the polarization of YouTube videos, which show that polarizing videos tend to trigger more diverse user-rating behaviors on comments and video. For identifying polarizing videos, we compute the difference of video and comments user-ratings. Thus, we compute the difference between the numbers of thumbs up (t_u) and thumbs down (t_d) as: $polarization = 1 - |(t_u - t_d)/(t_u + t_d)|$ for each video in our dataset[10]. Using this method our polarization range is between $[0, 1]$. For polarization range we derive 10 bins (such as 0-0.1). Comments on videos are assigned to a particular bin depending on the polarization topic of the related video. Then we estimate the prevalence of useful comments for each set related to each bin by using the usefulness classifiers and the Bayesian Prevalence Model.

The result of the relationship between the prevalence of useful comments and the polarization of topics of media objects is shown in Figure 3. We find the prevalence of useful comments decreases when the polarization of topics is higher. Furthermore, we inspected the coefficients of a linear regression model between the prevalence of useful comments on each video and polarization degree of the video. The coefficient rank (C= -3.362 p <0.01) indicates that the

[10]This experiment was conducted only on a YouTube set and not on a Flickr set, because Flickr photos do not have any rating.

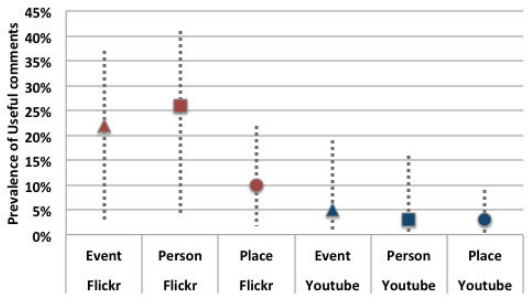

Figure 4: Different platforms (Flickr and YouTube) and topics lead to different usefulness prevalence.

polarization degree of topics has a negative correlation with the prevalence of usefulness. These results also support our findings regarding the time period effect of topics. The usefulness prevalences of some earlier periods (such as 1920s) are lower compared to those whose temporal dimension is later. This is because in these periods the selected topics are more polarized.

Influence of Entity Types of Topics on Usefulness Prevalence. Our result (shown in Figure 4) demonstrates that different platforms (Flickr and YouTube) lead to different usefulness prevalences. For all entity types of topics (place, person, and event), the usefulness prevalence of the Flicker platform is higher than that of the YouTube platform. Furthermore, Figure 4 demonstrates that the topic of the media object (event, place, person) leads to different usefulness prevalences. We get the lowest prevalence of useful comments for topics related to place for both platforms.

For YouTube, topics relating to person have a lower rate of comments than topics related to event. These results concur with our findings in the previous section that the most emotional topic is related to person and the less emotional a comment is, the more useful it is. In contrast, the topics relating to event have the highest rate of useful comments. Events may allow people to give more information about actual places, persons, and happenings. In this way, place and person topics are connected and consequently more information may be given. Contrary to what we expected, the rating results related to the different entity types of topics for Flickr are not similar to the prevalence results for YouTube. For Flickr, the highest prevalence for the three topics, person, place and event, is for person. For topics related to person on Flickr, we recognize that the time periods of many topics of selected photos are earlier compared to the time periods of topics of selected videos related to person for YouTube in our dataset. This is in line with our finding with regard to the effect of time period of topics on usefulness prevalence.

Discussion and Future Work

We conducted an analysis of user-generated comments on different social media platforms (Flickr and YouTube) to shed some light on the properties and prevalence of useful comments. The results of our analysis of three different sets of features — TL (text statistics and syntactic), ST (semantic and topical), and AS (user and social) — show that a few relatively straightforward features can be used to characterize and infer the usefulness of comments. It is interesting to note that many text features, while being positively aligned with usefulness inference, are not among the most important features. However, semantic and topical features play important roles. These results suggest that comments that contain a higher number of references, a higher number of named entities, fewer self-references and less affective language (lower sentiment polarity, lower subjectivity tone, swear score, etc.) are more likely to be inferred as USEFUL. An analysis of the usage of different terms shows that *insight* and *tentative* terms indicate a positive correlation with usefulness, while *certainty* terms do not. The analysis of features related to users suggests that by leveraging users' previous activities, we may be able to increase the likelihood of inferring the usefulness of a comment. This further suggests that users who mostly comment to converse and to describe their personal experiences (higher self-reference score) do not write useful comments. Moreover, users with a lower topical focus may write a comment with a lower level of focus about the specific topic, and, therefore, their comments are likely to be NOT USEFUL.

Another analysis of the important features among different entity types of topics (place, person, and event) indicates that when inferring the usefulness of comments, the influence of features varies slightly according to the topic areas of media objects. Major differences appear among the psychological characteristics of the content. Users express more emotion and may use more offensive language when writing comments about topics related to person and event. Such comments are more likely to be inferred as NOT USEFUL. Therefore, if prior to inferring usefulness we are able to determine the topic area of a media object, this helps in the classification of useful comments with greater accuracy. Thus, for a more accurate classification of useful comments, a classification model should be trained that takes into account the topic of media objects and the platform's commenting culture.

With regard to the analysis of the prevalence of useful comments, our findings indicate that prevalence is influenced by the commenting culture of platforms as well as the different dimensions of topics of media objects. The time period of topics has slight influence on the usefulness prevalence. The nearer the time period of a topic is to the present time, the lower the prevalence of useful comments is. Moreover, the polarization of topics has a negative contribution to the prevalence of usefulness. This means that for highly polarized topics the prevalence of useful comments decreases. Finally, we find that different platforms (Flickr and YouTube) lead to different prevalences of useful comments. For all entity types of topics (place, person, and event), the prevalence of useful comments on Flicker is higher than that of YouTube, which contains many more non-useful comments.

We also believe that topics related to a person's values and ideologies (countries, communism, capitalism, religion, nationalism, etc.) and other dimensions of topics of media objects (such as popularity and relevancy to present time) might influence the prevalence of useful comments. Furthermore, personal features (such as sex, location, education) of users may play an important role in the classification pro-

cess. Therefore, we will explore in future work these personal features and dimensions of topics on the prediction and prevalence of useful comments.

Acknowledgments

This work was supported in part by National Science Foundation Grants IIS-1111176 and IIS-0968450, and by a gift from Boeing.

References

Agichtein, E.; Castillo, C.; Donato, D.; Gionis, A.; Mishne, G.; Agichtein, E.; Castillo, C.; Donato, D.; Gionis, A.; and Mishne, G. 2008. Finding high-quality content in social media with an application to community-based question answering. In *Proceedings of WSDM*.

Ames, M., and Naaman, M. 2007. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07.

Becker, H.; Iter, D.; Naaman, M.; and Gravano, L. 2012. Identifying content for planned events across social media sites. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*

Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. In *the 20th international conference*, WWW.

Danescu-Niculescu-Mizil, C.; Kossinets, G.; Kleinberg, J.; and Lee, L. 2009. How opinions are received by online communities: a case study on amazon.com helpfulness votes. In *Proceedings of the 18th international conference on World wide web*, WWW '09.

Diakopoulos, N.; De Choudhury, M.; and Naaman, M. 2012. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, CHI '12. ACM.

Ghose, A., and Ipeirotis, P. G. 2007. Designing novel review ranking systems: predicting the usefulness and impact of reviews. In *ICEC '07: Proceedings of the ninth international conference on Electronic commerce*.

Gunning, R. 1952. *The Technique of Clear Writing*. McGraw-Hill, New York.

Harper, F. M.; Moy, D.; and Konstan, J. A. 2009. Facts or friends distinguishing informational and conversational questions in social q&a sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM.

Hsu, C.-F.; Khabiri, E.; and Caverlee, J. 2009. Ranking comments on the social web. In *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, CSE '09, 90–97. Washington, DC, USA: IEEE Computer Society.

Kennedy, L.; Naaman, M.; Ahern, S.; Nair, R.; and Rattenbury, T. 2007. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*.

Kim, S.-M.; Pantel, P.; Chklovski, T.; and Pennacchiotti, M. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06.

Lampe, C., and Resnick, P. 2004. Slash(dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04.

Liu, J.; Cao, Y.; Lin, C. Y.; Huang, Y.; and Zhou, M. 2007. Low-Quality Product Review Detection in Opinion Summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Liu, Y.; Bian, J.; and Agichtein, E. 2008. Predicting information seeker satisfaction in community question answering. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM.

Lu, Y.; Tsaparas, P.; Ntoulas, A.; and Polanyi, L. 2010. Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web*, WWW '10.

Momeni, E., and Sageder, G. 2013. An empirical analysis of characteristics of useful comments in social media. In *Proceedings of the ACM Web Science*, WebSci2013.

Ott, M.; Cardie, C.; and Hancock, J. 2012. Estimating the prevalence of deception in online review communities. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12. New York, NY, USA: ACM.

Siersdorfer, S.; Chelaru, S.; Nejdl, W.; and San Pedro, J. 2010. How useful are your comments?: analyzing and predicting youtube comments and comment ratings. In *Proceedings of the 19th international conference on World wide web*, WWW '10. ACM.

Sigurbjörnsson, B., and van Zwol, R. 2008. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08. ACM.

Tausczik, Y. R., and Pennebaker, J. W. 2010. The psychological meaning of words: Liwc and computerized text analysis methods.

Wagner, C.; Rowe, M.; Strohmaier, M.; and Alani, H. 2012. What catches your attention? an empirical study of attention patterns in community forums. In *ICWSM*.

Weinberger, K. Q.; Slaney, M.; and Van Zwol, R. 2008. Resolving tag ambiguity. In *Proceedings of the 16th ACM international conference on Multimedia*, MM '08. ACM.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*.